DOCUMENT RESUME

ED 365 107 FL 021 654

AUTHOR Barnwell, David Patrick
TITLE We Need an Empirical Basis for Evaluating College Foreign Language Teaching.
PUB DATE 93
NOTE 28p.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; College Faculty; *College Instruction; Evaluation Criteria; Evaluation Methods; *Faculty Evaluation; Higher Education; *Language Teachers; *Second Language Instruction; Teacher Effectiveness; Teaching Assistants; Test Reliability; Test Validity

ABSTRACT

College faculty are unlike other teachers in that they are not trained to teach and have little supervision. Effectiveness of language instruction is more difficult than most to evaluate. Contemporary models of supervision tend to emphasize training, not evaluation, even for teaching assistants. Some see teaching as an art, not susceptible to empirical study. Much current second language teaching research is insubstantial. Existing approaches to classroom observation include rating scales, checklists, time-based or event-based tabulations, and ethnographic methods. Some methods are high-inference and some low-inference, but results most often involve a high degree of inference. More attention should be given to the observation/evaluation instruments used. Minimum requirements for a useful instrument include practicality, reliability, and validity. The profession should develop profession-wide rather than institution-based instruments reflecting relevant theory and practice. There exist models for development of such a measure, and it is possible to make one that is dynamic, capable of adaptation to new research evidence. Issues that might be addressed in creation of an empirically-based teaching evaluation instrument include quantity and quality of teacher talk, comprehensible input, formal grammar practice, error correction, teacher-student interaction, and teacher experience. Contains 36 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 365 107

WE NEED AN EMPIRICAL BASIS FOR EVALUATING COLLEGE FOREIGN LANGUAGE TEACHING

David Patrick Barnwell

FL021654

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Barnwell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

David Barnwell

WE NEED AN EMPIRICAL BASIS FOR EVALUATING COLLEGE FOREIGN LANGUAGE TEACHING

BACKGROUND

In the last few years we have seen the development of a body of literature on Language Course Supervision. Many interesting contributions have been made to the literature in this area, especially in regard to the application of new technologies to the training of foreign language teachers. Yet there remains a curious gap in the burgeoning professional literature of language course supervision. There is a lot on how to train teachers, but almost nothing on how to judge the outcomes of that training. Many suggestions are offered as to what will improve instructors' performance, but few ideas as to how to assess whether that improvement has been attained. Do we have any tenable criteria of teaching effectiveness, both to be used by those who observe classes and to be set forward as goals for instructors ? Is class observation as currently carried out in the foreign language classroom truly practiced on a professional basis ? It will be suggested here that the profession now needs to consider the methodology to be used in evaluating the teaching of foreign languages.

How shall we rate teachers' effectiveness ? What about using student achievement as a benchmark ? Europe used to have the system of "payment by results", in which teachers' salaries were determined by how well their students

performed on outside tests. To judge teachers on the basis of their students' achievements has the advantage of apparent objectivity. Unfortunately, foreign language testing instruments have tended to lag behind the state of the art in language teaching. Thus a teacher who is preparing students for a standardized or national examination will very often have to use a methodology that is not representative of current thinking in foreign language teaching theory. Further, to base judgments of teaching on the performance of heterogeneous student groups raises so many issues of fairness that it is inconceivable in today's climate.

There are also other student-based indices, for example student evaluations. Logically students in a particular course should be good evaluators of both instructor and instructional materials. There is no evidence that students are unfair in their evaluations, certainly when these are considered en masse. Yet, though recognizing the importance of student input, it is doubtful that we would cede to students the right to hire or fire their teachers. For one thing, students' goals in taking a course can differ substantially from the aims of the institution or instructor offering that course. Further, since they are by definition in very partial command of the subject matter, students cannot have a total view of the instructional sequence and how each element may fit into it. In some institutions ratings provided by student evaluations are used in tenure and promotion decisions. Whether a valid belief or not, many teachers' grading policies would be affected by the

suspicion that the grades they awarded could influence the evaluations they received. In addition, the foreign language teacher is subject to a little bias specific to the system itself. Foreign language teachers teach a disproportionately high number of required courses, and these are at elementary and intermediate levels. It has been found that the lower the level of study, the lower courses and teachers tend to be rated. Further, courses taken as a requirement tend to be rated lower than courses taken optionally (Pennington & Young 1989). This may tend to shade foreign language teaching evaluations a little downward when compared to other faculty.

In any case it is certainly true for all faculty that in no other walk of life are judges called on to administer an evaluative instrument in which they have no training, and whose uses they may not even understand. In no other area of life would anonymous opinions be given such credence in a decision of major importance. Yet this is the case once student evaluations are used to judge teacher performance. There might even be constitutional questions here--with regard to the right to identify one's accuser--were a fired faculty member to seek recourse through the courts against decisions made on the basis of student evaluations.

Perhaps there should be no evaluation at all. This is after all the case with the teaching performance of many of the tenured faculty at our universities. Yet the context in which foreign languages are taught in the larger American universities does not permit evaluation to be abandoned. In the first place, assessment of the performance of instructors is a prerequisite to

guiding or helping them to improve their teaching. Secondly, as representative of the university, and acting for those sometimes hundreds of undergraduate students taking language courses, it is incumbent on a department to define standards of teaching and see that Teaching Assistants and part-time instructors make progress towards reaching them. In extreme cases, an instructor who fails to make some approximation, however vague, to these standards may have to find some other way of earning a living or financing graduate work.

As those who supervise and evaluate are dealing with people's careers and livelihoods it is hardly surprising that there exists a body of writing on the legal aspects of classroom observation and teacher evaluation. Like most of the literature in this area, the focus is on high school and elementary levels. Broadly, Morris and Curtis (1983) and Wise and others (1983) report that in high-school systems, just as in any other job-evaluation procedure, the courts have declared that there must be predetermined standards which are clearly defined and observable, and can be shown to be job-relevant. These must be public, and readily available to the evaluated person. If the evaluation shows the individual not in compliance with these standards, s/he must be given an opportunity to correct any deficiencies. Actually there have been surprisingly few litigations arising out of adverse evaluations of high-school teachers, but this number is likely to grow. More and more states are laying down guidelines for teaching evaluation (Lewis 1985).

An awareness of legal precedent from the high schools is useful in reminding us what neutral parties have judged to be fair standards to aim for in professional evaluation. Although recourse to the courts has spread in academe during the past few years, especially around tenure decisions, there has not been much litigation concerning those part-time faculty and graduate students who teach foreign languages in college or university. At almost every institution such personnel are part-time workers, non-unionized and comparatively unorganized. They are transient workers and their dual role of student-workers does not yield any great bargaining power. In foreign languages--as increasingly in many other areas--quite a few of the Teaching Assistants are non-citizens of the U.S., academic braceros or guest workers who are not always in a position to raise objections to perceived anomalies in evaluation. Thus the legal questions, though we should surely be aware of them, have so far proved tangential to the work of those who direct language courses. But because the individual Teaching Assistant is seldom represented by any professional body it is all the more incumbent on departments, more specifically on Language Course Directors, to see to it that their evaluation procedures are fair and valid. Clearly, just because Course Directors are unlikely to have to offer legal justification of their procedures they are not absolved of their ethical and professional obligations. As professionals, they should be able to stand over the validity of what they do, especially when it can profoundly affect others. Happily, in the great majority of cases, the work of the Course Director is purely advisory--there is no question of an instructor's career

being in jeopardy. But even in these less controversial instances, the need for good observation and evaluation persists. It is rather striking to realize that a large number of those of our Teaching Assistants who go on to graduate and get university positions will never be observed again in their lifetimes. All the more reason to hope that our observation and evaluation will have something valid to say to them now.

The Director/Coordinator is typically both supervisor and evaluator. In other words, not only does s/he work with the instructors to improve standards of teaching, s/he must also determine to what extent an individual has achieved this goal. There is scope for a certain tension between these two facets of the coordinating duties. Think of the driving test--would it be valid for the same person to act as both driving instructor and tester ? Language Directors are called on to judge the very people they are training, even though such a judgment is in part a verdict on how effective that very training has been. It is a somewhat inbred system, which would benefit from more professional accountability on the part of the Director. The position parallels that of the class teacher, who is also is called on to grade/evaluate those whom he has been training, namely his own students. Surely if we tell our instructors that there are certain minimum requirements that classroom tests should meet, so also should we be confident that our own evaluative instruments meet these criteria in validity and reliability. Hence it is time to reflect upon the methodology of how we evaluate.

OBSTACLES TO OVERCOME

Contemporary models of supervision tend to emphasize the training aspect--the supervisory dimension, rather than the evaluative. In fact, it sometimes appears that Language Course Directors are a trifle apologetic for the evaluative component of their work, preferring instead to focus on the apparently more pleasant and more positive aspects of working cooperatively to heighten instructors' self-awareness and to help them analyze their own teaching. In contrast, evaluating Teaching Assistants, maybe even to the extent of labeling an individual as incompetent, can seem a little old-fashioned and out of tune with the individuality and creativity we seek to foster in our classes. Though less so in the United States than in other countries, even here there is a long tradition that the university classroom is hermetic, that "it is really ungentlemanly to peek" (Hayes et al., 1967 p.8).

There are also other players--often powerful ones--on the departmental stage, in the form of colleagues who sometimes have their own agenda for graduate students. The Course Director-instructor relationship is but one line on a triangle; both have another line to the department as a whole. There appear to be still some cases of professors who view graduate students' teaching as a diversion from the students' true vocation of populating graduate courses. Rarely is a Teaching Assistant's effectiveness as a language teacher the only factor taken into account in assessing his progress in a department. Issues such as these tend to provide an unwritten agenda for those who carry out language teaching evaluations.

Language acquisition research has grown to constitute an unwieldy and confusing body of disparate materials. Universal and easily-applicable generalizations as to what constitutes good teaching do not leap from the pages of the books and articles we read. The profession seems to be aware of its susceptibility to bandwagons, even to the extent that no certainties are tenable any more. As Allwright (1983, p.199) puts it, the last decade or two has seen "the retreat from prescription to description, and from technique to process". No longer are laws for teaching behavior laid down, be they at the level of grand method or mere humble classroom technique. Today's research instead sets out to describe the processes at work in the classroom and outside.

Thus some teachers--even some language course coordinators--appear to see teaching as an art; you either can do it or you can't. Humans are so complicated, it is argued, that one cannot draw up generalizations about anything. Good teaching then is mysterious, even indefinable. All very well, but a couple of centuries ago, as is pointed out in Dunkin and Biddle (1974, p.8) surgery too was considered an art. Yet would we prefer to be operated on by a surgeon who is aware of all the latest empirical data and who seeks to base his procedures on the scientific method ? Or by someone who can't explain how he does it, why it works when it does, or why it didn't when it didn't ?

There is the further view that we do not know enough about teaching to be able to stand over anything we say. What

works for you may not work for me; your guess is as good as mine. Certainly it is true that it would be impossible to prove an inevitable cause-effect connection between particular teaching behaviors and good teaching outcomes, however measured. But as Gage (1978, p.234) notes, the same could be said about the cigarettes/cancer research. One cannot predict that any particular smoker will develop lung cancer, nor that any particular non-smoker will not. But does that preclude us from having confidence in the research that shows a relationship between the two ? If one's child asked whether he should take up smoking, would the reply be: "Well, there are so many variables in the research that I really can't give any hard and fast answer" ?

This is by no means to make unthinking obeisance to research as if it could provide all the answers. A lot of what has been published in classroom second language acquisition research is tentative, even sometimes trivial. Undoubtedly in ten or twenty years new questions will have been posed and perhaps some of today's answered. But those who direct language courses work every day at the interface between theory and practice, and cannot afford to wait until a universal theory has been worked out. As part of the drive towards full professionalization of the work that such faculty do, we need to explicitly incorporate a research-based component, rooted in what is known today rather than what may be found out in a few decades.

OBSERVATIONAL AND EVALUATIONAL FORMATS

Instruments for observing and evaluating teaching have quite a long tradition (Gilmore 1927, Puckett 1928). An enormous variety of procedures has been developed, be they for evaluative or research purposes. In 1948 Barr reviewed 209 scales for rating teachers. Domas and Tiedemann's bibliography on the subject cited 663 articles or books. More recently, Borich and Madden (1977) transcribe over 200 published instruments. It is not likely that many Language Course Directors are familiar with the literature in this field. Given the heterogeneous academic and professional background of such individuals, it is more probable that very few have received formal training in any observation technique. An unmoveable obstacle to the effort to professionalize the work of Language Course Directors is the fact that no foreign language department offers even partial preparation in the techniques of supervision as part of its graduate program. Even those Course Directors who have substantial training in second language acquisition, pedagogy and applied linguistics, are unlikely to have made the explicit study of the process of supervision which might foster the study of observational techniques. In a sense no Language Course Director has been comprehensively trained for the work s/he does; some are just less inadequately prepared than others. What we need is a theory of classroom observation and evaluation.

In an effort to bring the two worlds together, that of the foreign languages department and that of the education department, it may be worthwhile briefly to review some existing

approaches to classroom observation. Generally, one can discern four groups:

1/ Rating Scales

These measure the extent or frequency of some particular manifestations. Many commentators stress the potential for unreliability in rating scales, and counsel that all who are to use them be trained. Of course in the language department it is often the Director who drew up the rating scale in the first place. Psychometrically, there is room for argument as to the validity of using a scale invented by the user. At the least, it is not conducive to the maintenance of objective and profession-wide standards. Nevertheless, rating scales are in common use in foreign language departments. Omaggio (1986, p.471-2) offers a sample. Here the instructor is rated on a four-point scale, ranging from (1)-needs much improvement, to (4)-outstanding. Data are elicited by questions ("Did the instructor seem to have planned the day's lesson to include communicative practices?") or by statements ("Entire class was involved in the lesson"). Omaggio's instrument is one of the most recently published, and shows an awareness of developments in methodology in the last decade or so. It provides a a useful point of departure in a discussion of foreign language teaching evaluation and observation. Even so, it begs its own questions--what practices are communicative ?

2/ Checklist

As in the previous case, the design of the checklist is of great importance, since it a priori sets out what should be looked for, what is "good" or "bad". This format seeks to be

more objective and reliable than the rating scale. Rather than give a rating, the observer checks whether or not a particular behavior or quality has been shown. It is thus more reliable, but less flexible and subtle. A brief passage from Morin and Lemlech's (not specific to foreign language) checklist (p.87) will give a flavor:

Teacher demonstrated awareness of students' needs	YES	NO
- monitored students' work	YES	NO
- provided feedback	YES	NO
- facilitated participation	YES	NO
- facilitated thinking	YES	NO
- questioned students	YES	NO
- listened/observed/took notes	YES	NO

However, the checklist's promised objectivity and reliability does not inevitably materialize. In the study by Morin and Lemlech, in about half of 45 lessons observed there was clear disagreement between the observers. For instance, a category such as "Teacher used a variety of techniques" elicited great disagreement. Some observers said she did, others said she didn't.

A checklist can be as comprehensive or as exclusive as we wish. It might contain only one category, were that category deemed indispensable to good teaching. Harrington (1955) tallied only teacher smiles and ignored everything else. Is there any single category that could be considered indispensable for the foreign language teacher ?

3/ Tabulations, time-based or event-based.

These are probably the most widely-used instruments in research on teaching. A version that is still influential is Flanders' Interaction Analysis, which was adapted specifically

for the language classroom by Moskowitz (1971). Moskovitz's FLint seeks to track teacher talk and student talk. It consists of twenty categories, twelve of these for Teacher Talk, three for Student Talk, and five for such things as laughter, silence etc. The first six categories of Teacher Talk are as follow:

1. Deals with feelings.
2. Praises or encourages.
2a. Jokes.
3. Uses ideas of students.
3a. Repeats student response verbatim.
4. Asks questions.

The observer maps what is occurring in the classroom through making a tally based on categories such as those above. In the case of FLint, these tallies are made at three-second intervals. The resulting patterns that describe what has gone on will, it is hoped, be comprehensive and analyzable, and constitute a valid record of what transpired in class. Discourse instruments of this type require a good deal of training for those who use them, and can pose almost intractable difficulties in coming to a common understanding of how particular events in the classroom are to be categorized. Their terminology can be somewhat opaque, tending towards jargon. Looking through Borich and Madden (1977, p.173) one comes across a system in which a teacher's "uh uh" is recorded as "Minimal Reinforcement". "Non-verbal affiliation", we are told here, includes physical contact, such as a teacher's putting an arm around a student. More importantly, these instruments were designed before today's fondness for group or paired work--they are very much oriented to the teacher-student axis rather than student-student. So Category 11 of Moskovitz's FLint system is

CONFUSION, WORK-ORIENTED: More than one person at a time talking, so the interaction cannot be recorded.

Such a category would have to be revised for the classrooms of the 1990s.

4/ Ethnographic

This has been widely used in first language settings. It is concerned more with observation than with evaluation. The observer seeks to understand a situation on its own terms, rather than force it into an external or a priori construct or value judgment. The ethnographer rejects the role of disinterested outsider and willingly participates in what is being observed. Watson-Gegeo (1988, p.583) describes the approach thus: "One of the hallmarks of ethnographic method is intensive, detailed observation over a long period of time. Ideally, an ethnographer observing a university-level ESL class, for example, would observe all class meetings for the entire semester, conduct interviews with a sample of the students and the teacher, and observe the students in other settings." Procedures associated with the ethnographic tradition are so varied and eclectic that no proper sample can be given in the space available here--see Cazden et al. (1980) or the articles collected in Green and Wallat (1981). The Clinical Supervision movement, which has been influential in general high school supervision for the past three decades or so, shares many of ethnography's collaborative and non-judgmental instincts (Acheson and Gall, 1987).

Faculty who work in the training and evaluation of part-time instructors would do well to familiarize themselves with the rich and humanistic philosophy embodied in this tradition. As a caveat, however, they might consider to what extent such models are compatible with the Course Director-Teaching Assistant relationship. For one thing, they are hardly practical within the constraints of directing a large number of instructors. One cannot carry out ethnography or Clinical Supervision based on a once-a-semester class observation. Further, Clinical Supervision assumes that the supervised teacher enjoys full professional status--the process is one of interaction between equals. This does not reflect the reality of the situation in which Course Directors work.

DEGREE OF INFERENCE

Rosenshine (1970) distinguishes between high-inference and low-inference procedures. Low inference systems require the observer to make few inferences or interpretations about what he sees. As one would expect, they are more reliable than high-inference procedures, since there is less contamination from observer variables. However, there are probably lots of things which cannot be spelled out in low-inference terms e.g., the clarity of an explanation. Clarity is more a function of audience variables than it is of the intrinsic quality of the explanation. It is improbable that it could be objectified by a measure such as length of explanation, number of examples or questions asked about the explanation, success of students on subsequent task or whatever. Low-inference can also somehow

miss the essence of what is going on. Compare for instance the same behavior as described first by high-inference and then by low: (Yuzdepski 1985a, p.67)

> The teacher appeared warm and convivial as a thorough set of instructions was presented to the class. The students were eager to learn and listened attentively.
>
> VERSUS
>
> Teacher delivered verbal instructions to the students. Students silent.

It is not the purpose here to argue for the relative superiority of any one of the four formats that have been sketched. A comprehensive discussion of observation systems as specifically applied to the language classroom is provided by Chaudron (1987) and Allwright (1983). Both these treatments of the topic, though fine, are somewhat focused on the ESL classroom--the present proposal is that the discussion be taken up by the foreign language teaching profession. In the end, whether the observation procedure is high or low inference, it will probably end up as high inference. Unlike the ethnographers, those who supervise language teachers do not have the luxury of being neutral about what they witness when they observe; for them observation and evaluation are symbiotic. A Course Director who went to a foreign language classroom with no preconceptions of what is good and bad teaching would be earning money under false pretenses. No Course Directors who have at all thought about language teaching are going to gather a set of data and leave them at that. They are going to ask "What does all this mean ?", and once they start doing so they are making inferences.

PROFESSIONAL, NOT PAROCHIAL OBSERVATIONAL PROCEDURES

As an aid and focus in carrying out observation and evaluation, many departments continue to use an Observation Sheet, evaluation schedule, visitor's form--call it what they will. This Class Observation Sheet is quite an important document. Not alone does it serve in the evaluation of particular individuals, it constitutes a general statement of the Department's views on foreign language teaching, and defines what the Department considers to be desirable in the classroom. As part of the task of inducting new instructors into the department, they are often issued with a copy of the Observation Sheet at the beginning of their teaching. They thus have an explicit set of desiderata to aspire to in their daily practice. In those institutions which offer a Teaching Assistant orientation or methods class, discussion of such an Observation Sheet would appear also to have its place.

It is time to devote more thought to the status of these evaluation instruments. Though in theory they ought to represent the department's distilled wisdom as to what constitutes good teaching, in practice they are more like the convention platforms of Democrats and Republicans--heatedly debated for a few days every few years, and subsequently quite marginal to what really goes on. Let us dust off our evaluation instruments, update them in the light of the best of current research, and then stand over them with confidence.

There would seem to be several minimum requirements in setting our evaluation mechanisms on a proper professional basis.

Viewing the observation and evaluation as a kind of test permits us to borrow many of the basic considerations in testing theory. Three immediately come to mind:

1/ The instrument should be useful and practical. It should not be excessively cumbersome or require extensive training in its use. It should avoid jargon.

2/ The instrument should be reliable--or more accurately it should be capable of being used reliably, since essentially it is the observer rather than the instrument who is carrying out the observation. Similar performance should receive similar treatment on the scale. Reliability should extend across raters--the data should not be a function of the observer's particular characteristics. Of course the prescriptive purposes of the instrument would lead us to welcome a certain lack of coherence between different administrations (test-retest), since one would hope that weaker performances would improve.

3/ The instrument should be valid. Construct and content validity would be established on the basis of the instrument's congruence with what research and professional literature have established about classroom second language learning. Every element in the instrument would be subject to scrutiny in the light of the current state of the literature.

Those who are charged with working with part-time faculty would do well to consider the desirability of more cooperation and coordination in the evaluative and observational techniques they use. We should seek to develop profession-wide rather than parochial instruments, incorporating not just individual intuitions and institutions but also basing ourselves on a pan-professional awareness of relevant theory and practice. Models exist for how to go about this. Foster (1983) reports on how 199 indicators of teaching effectiveness were boiled down to a more manageable list for use in teacher assessment. These 199 indicators came from the researchers' analysis of the relevant literature, and they were then in turn rated for importance by a

panel of educators. It was found that the essentials of the 199 indicators could be embodied in 47 components of teaching effectiveness. This research was oriented towards global teaching ability rather than any specific content area.

Foster's account shows that it was quite a complicated task, as it would be in the case of foreign language teaching. It is nonetheless a very feasible one. The first step for foreign language might be a polling of readers of professional journals, in which respondents would be asked to list those features of teacher performance that they believed to be of importance in the foreign language classroom. An important innovation here might be to call for each assertion to be supported by reference to the literature of pedagogy or second language acquisition. One might be surprised by the degree of coherence in the responses. Over twenty years ago, Hayes, Lambert and Tucker (1967) carried out a similar survey for foreign language and found "marked agreement as to what is important and what is not". The disparate responses could then be reduced to a manageable list that included as much of the common criteria as possible, and was couched in terms of one of the models mentioned earlier in the discussion on formats of evaluation. At this stage, too, a strong empirical and bibliographical component would be desirable, in an attempt to root the instrument in the tradition of research rather than of intuition. The list might either be pared to a minimum, to express a small set of bare essentials for good teaching, or larger and looser, to allow for individual variation among teachers and programs. The instrument would not be ossified,

but rather should be thought of as dynamic, capable of changing in light of new evidence from research. Generally, the work would benefit from the input of our ESL colleagues, who have for a few years now been attempting to devise measurements of the teaching effectiveness of non-English speaking Teaching Assistants at American universities (Pennington and Young 1989).

The pilot instrument would undoubtedly be subject to trialling and validation, both in actual classroom use and in terms of how it reflected the best of the literature on classroom second language acquisition. This very process would be certain to promote useful dialog within the profession, as the validity of each person's conceptions about good teaching was exposed to the critique of colleagues. In seeking to clarify specific classrooms behaviors or skills, methodologists might be prompted to consider further how such desirable teaching practices may be fostered or acquired. It would be a learning experience for all students of pedagogy and second language acquisition to have to cite references in the literature in support of the points they argued for.

What kinds of issues might be addressed in the creation of an empirically-based teaching evaluation instrument ? Let us take a few examples. Isn't it time to reconsider the formerly somewhat negative categorization of quantity of teacher talk, given the research on the utility of a "Silent Period" in foreign language acquisition (Postovsky 1970) and the current popularity of the notion of "comprehensible input" (Krashen 1982). What is the role of formal practice in grammar (Ellis 1989) ? Does

research offer any suggestions about the quality or kind of teacher talk (Henzl 1979) ? What about quantity or quality of student talk (Brock 1986) ? Then there are correction techniques--what to correct, when to correct, how to correct, who should correct (Hendrickson 1978, Chenoweth et al., 1983). Group work, student-student interaction--can we come up with any empirically-based generalizations ? (Pica & Doughty 1985, Long & Porter 1985). Further questions remain, somewhat more specific to the work of Course Directors. Should a set of desiderata for Teaching Assistants differ from a similar set for teachers at other ranks--less recognition for innovation and a greater stress on following a common program, perhaps ? To what extent should an instrument seek to be globally valid, and to what extent should it make allowance for specific factors such as goal and level of course, composition of student body, availability of technical resources etc. ? Are there specific needs for the advanced class, perhaps overlooked by methodologists' tendency towards preoccupation with elementary levels (Gutierrez 1990) ? Many other examples could be cited, of areas in which teaching evaluation instruments need to catch up with and synthesize the state of the art in such things as the teaching of culture, the use of realia and authentic materials, exploitation of newly-available technologies, and so forth.

If a teaching evaluation instrument is considered as a rating scale used in a test, and that is what it is, the instrument should be expected to bring with it some statistics on validity and reliability, as well as directions on how it is to

be administered. It seems a fair bet that no such information is available on any instrument in use in university foreign language departments at present.

The extent to which such a project as described here could be implemented would be indicative of whether or not classroom-based language learning research, a field which goes back at least to the great Modern Foreign Language Study of the 1920s, has as yet come of age. Even in the case that such a project did not come to full fruition, the discussion it generated would yet be of great benefit to the profession, since it would provide the bridge between theory and practice that many of us seek. This may well be the direction that research will take in the later 1990s--away from the elaboration of theories of second language acquisition, and towards a concentration on what really goes on in classrooms and other learning settings.

NOTES

1. Yuzdepski (1985) provides a model for what a bibliographically-based instrument might look like.

REFERENCES

Acheson, Keith, and Meredith Gall. 1987. Techniques in the clinical supervision of teachers. New York: Longman.

Allwright, Dick. 1983. Classroom-centered research on language teaching and learning. TESOL Quarterly, 17,2: 191-204.

Barr, A.S. 1948. The measurement and prediction of teaching efficiency. Journal of Experimental Education, 16: 203-83.

Borich, Gary, and Susan K. Madden. 1977. Evaluating classroom instruction: a sourcebook of instruments. Rowley, Mass:Addison-Wesley.

Brock, Cynthia. 1986. The effect of referential questions on ESL classroom discourse. TESOL Quarterly, 20,1: 47-59.

Cazden, Courtney, Robert Carrasco, A.A. Maldonado and F Erickson. 1980. The contribution of ethnographic research to bilingual bicultural education. In James Alatis ed., Current issues in bilingual education, 64-80. Washington, D.C: Georgetown University Press.

Chaudron, Craig. 1987. Second language classrooms. New York:Cambridge Univ. Press.

Chenoweth, N Ann, Richard R. Day and Stuart Luppescu. 1983. Attitudes and preferences of ESL students to error correction. Studies in Second Language Acquisition 6,1: 79-87.

Domas, Simeon, and David Tiedeman. 1950. Teacher competence: an annotated bibliography. Journal of Experimental Education, 19: 101-218.

Dunkin, Michael and Bruce Biddle. 1974. The study of teaching. New York: Holt Rinehart & Winston.

Ellis, Rod. 1989. Are classroom and naturalistic acquisition the same ? A study of the classroom acquisition of German word order rules. Studies in Second Language Acquisition, 11, 305-28.

Foster, Clifford. 1984. Selection of evaluation criteria for the development of a teacher assessment system. ERIC EDRS ED 242719.

Gage, N.L. 1978. The yield of research on teaching. Phi Delta Kappan, 60,3: 229-35.

Gilmore, M.E. 1927. Judging and rating the teacher. Educational Review, 74: 269-72.

Green, Judith, and Cynthia Wallat. 1981. Ethnography and language in educational settings. Norwood, New Jersey: ABLEX.

Gutierrez, John. 1990. Overcoming anarchy in the advanced language class. ADFL Bulletin 21: 40-6.

Harrington, G.M. 1955. Smiling as a measure of teacher effectiveness. Journal of Educational Research, 48: 715-7.

Hayes, A.S., W.E. Lambert, and S.R. Tucker. 1967. The evaluation of foreign language teaching. Washington: Center for Applied Linguistics. ERIC EDRS ED 291250.

Hendrickson, James. 1978. Error correction in foreign language teaching: recent theory, research and practice. Modern Language Journal, 62: 387-98.

Henzl, Vera. 1979. Foreigner talk in the classroom. IRAL 17: 159-67.

Krashen, Stephen. 1982. Principles and practice in second language acquisition. Oxford: Pergamon Press.

Lewis, Anne. 1985. Evaluating education personnel. ERIC EDRS ED 212055.

Long, Michael, and P Porter. 1985. Groupwork, interlanguage talk, and second language acquisition. TESOL Quarterly, 19, 202-228.

Morin, Joy, and Johanna Lemlech. 1987. Supervising teachers and the University Supervisor's perceptions of teaching behaviors. Teacher Education Quarterly, 14,4: 84-94.

Morris, John E., and Fred Curtis. 1983. Legal issues relating to field-based experiences in teacher education. Journal of Teacher Education 34,2: 2-6.

Moskowitz, Gertrude. 1971. Interaction analysis--a new modern language for supervisors. Foreign Language Annals, 5:211-221.

Omaggio, Alice. 1986. Teaching language in context. Boston: Heinle & Heinle.

Pennington, Martha, and Aileen Young. 1989. Approaches to faculty evaluation for ESL. TESOL Quarterly, 23: 619-46.

Pica, Teresa, and Cathy Doughty. 1985. Input and interaction in the communicative language classroom. In S.M. Gass & C.G. Madden eds., Input and second language acquisition, 115-32.

Postovsky, V. 1970. The effect of delay in oral practice at the beginning of second language teaching. Ph.D. dissertation, UC Berkeley.

Puckett, R.C. 1928. Making supervision objective. School Review, 36: 209-12.

Rosenshine, B. 1970. The evaluation of instruction. Review of Educational Research, 40,2: 279-300

Watson-Gegeo, Karen. 1988. Ethnography in ESL; defining the essentials. TESOL Quarterly, 22: 575-92.

Wise, Arthur, Linda Darling-Hammond, & Sara Pease. 1983. Teacher evaluation in the organizational context. Review of Educational Research, 53,3: 285-327.

Yuzdepski, I. 1985a. Planning for an evaluation of teaching performance: manual of guidelines. Alberta Dept. of Education: Edmonton. ERIC ED 270399.

Yuzdepski, I. 1985b. Evaluation procedures: annotated bibliography and references. Alberta Dept. of Education: Edmonton. ERIC ED 270401.

-o-o-o-o-o-o-o-o-o-o-o-